Employees Dataset Analysis Using PySpark

Office Employees Data Analysis

Source: : https://github.com/PacktPublishing/50- Hours-of-Big-Data-PySpark-AWS-Scala-andScraping/tree/main/Part%203/Code/03- Spark%20DFs

Dataset: HR_Analytics

✉ Email: ankitabhamidimarri21804@gmail.com
☎ Phone : 6304450121
in LinkedIn : https://www.linkedin.com/in/ankita-bhamidimarri/

Bharadwaj Kollepara

Introduction

This presentation provides an analysis of the Office Employees data, covering:

- Structure and Uniqueness

- Departmental Distribution

- Attrition by Department

- Salary Insights

- Age Distribution

- Attrition Patterns

- Workforce Characteristics

Initial Analysis of the Dataset

1. Structure and Uniqueness

he dataset contains details of 1,470 employees with 7 columns: EmpID, Age, Department, JobRole, MonthlyIncome, Attrition, and PerformanceRating. Each employee is uniquely identified by EmpID; employee names, state/location, and bonus fields are not present. Data types are correctly inferred for analysis; explicit checks for missing values or duplicates were not performed in the notebook. Departments span three categories, capturing a compact workforce structure suitable for department- and role-level analytics.

2. Departmental Distribution

There are three departments represented, with Research & Development being the largest (961 employees), followed by Sales (446) and Human Resources (63). This R&D-heavy structure indicates a product/research-centric organization, while Sales forms a substantial go-to-market arm and HR remains lean. Such distribution informs workforce planning, managerial span of control, and support function resourcing, and it provides context for interpreting salary and performance differences across roles.

Initial Analysis of the Dataset

3. Attrition by Department

Attrition counts vary across departments, with higher absolute attrition visible in Research & Development simply due to its larger headcount, followed by Sales; Human Resources shows the fewest cases. While counts were visualized, attrition rates were not computed—normalize by department size for fair comparison in future work. This view helps flag potential retention hotspots and prioritizes tar

4. Salary Insights

MonthlyIncome is right-skewed with substantial high-end outliers, reaching ~20,000. Leadership roles earn the most on average (Managers ≈ 17,182; Research Directors ≈ 16,034), while individual-contributor roles such as Sales Representatives and Lab Technicians sit in lower bands (~2,600–3,300). Departmental ranges are broad—R&D: 1,009–19,999; Sales: 1,052–19,847; HR: 1,555–19,717—indicating wide intra-department dispersion. Use a histogram and boxplot to show distribution and outliers, and a bar chart for average salary by JobRole or Department.

Initial Analysis of the Dataset

5. Age Distribution

The workforce spans from at least age 18, with most employees concentrated in the late-20s to early-40s range. A meaningful senior segment exists (>45), for which a focused subset was created in the analysis. Use a histogram (with KDE) to show the overall shape and a boxplot to highlight any age outliers. If needed, annotate the >45 share to emphasize senior-talent presence.

6. Attrition Patterns

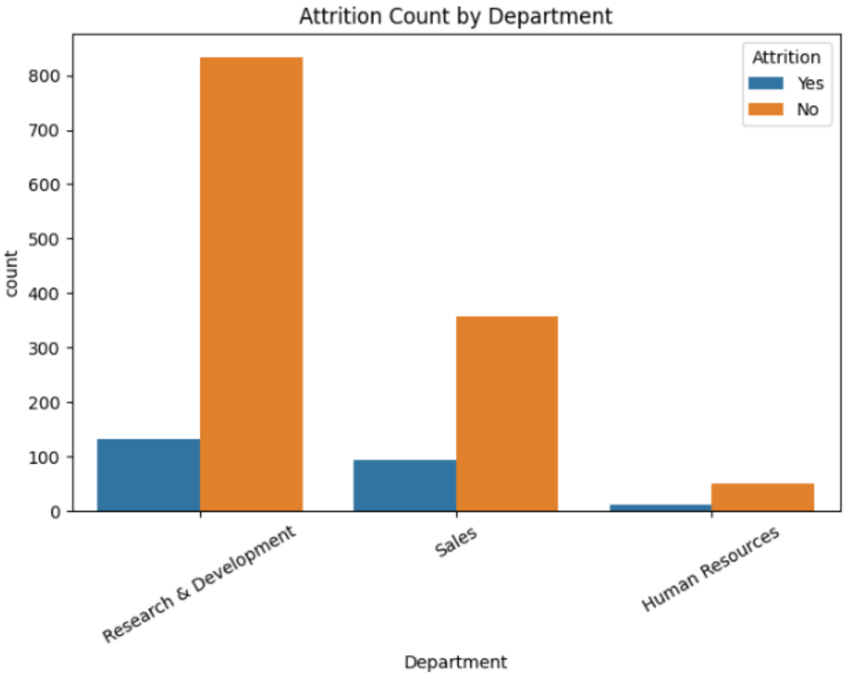
Attrition varies by department, with the highest number of cases in Research & Development, followed by Sales, and the fewest in Human Resources—largely mirroring each unit's headcount. Use a stacked bar chart of Attrition (Yes/No) by Department to visualize differences; for fair comparison, add a rate view by normalizing counts to department size.

Initial Analysis of the Dataset

8. Workforce Characteristics

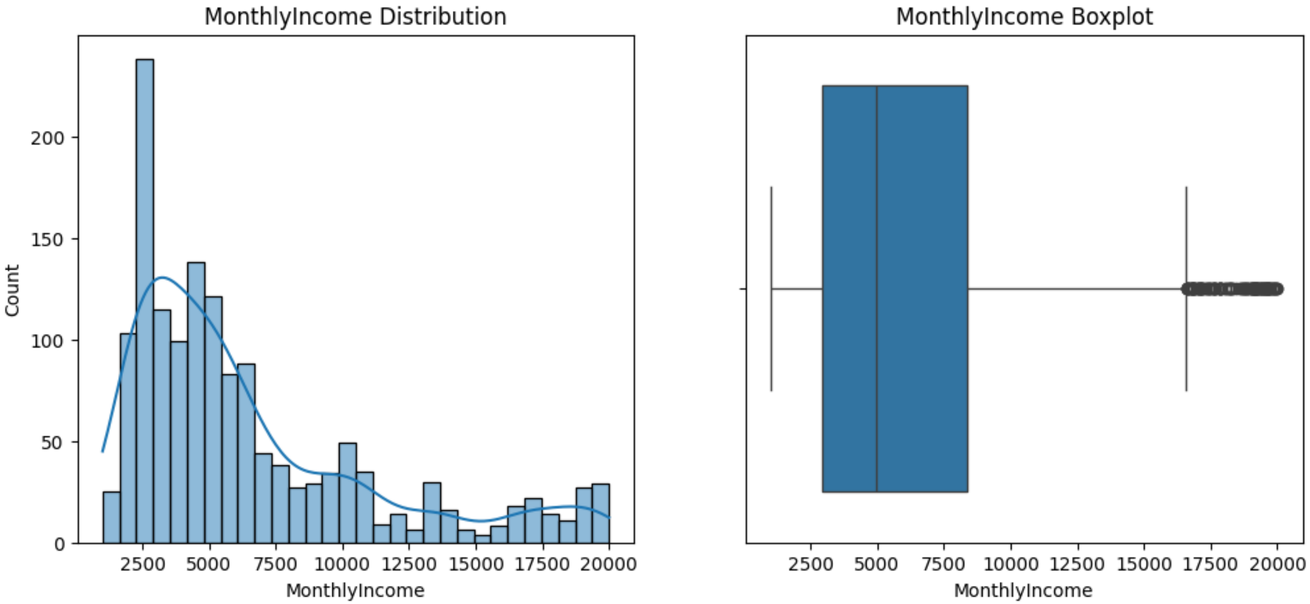
Overall, the company has a mid-sized workforce (1,470 employees) with an R&D-centric structure, a substantial Sales function, and a lean HR team. The age mix spans early-career to senior employees, with a notable >45 segment; performance ratings cluster around 3.1–3.2 across roles, indicating stable evaluations. Compensation is right-skewed with high-end outliers, and leadership roles earn the highest averages. The dataset’s well-typed schema enables reliable department/role analytics; adding location and bonus fields would broaden workforce insights.

Attrition count by Department



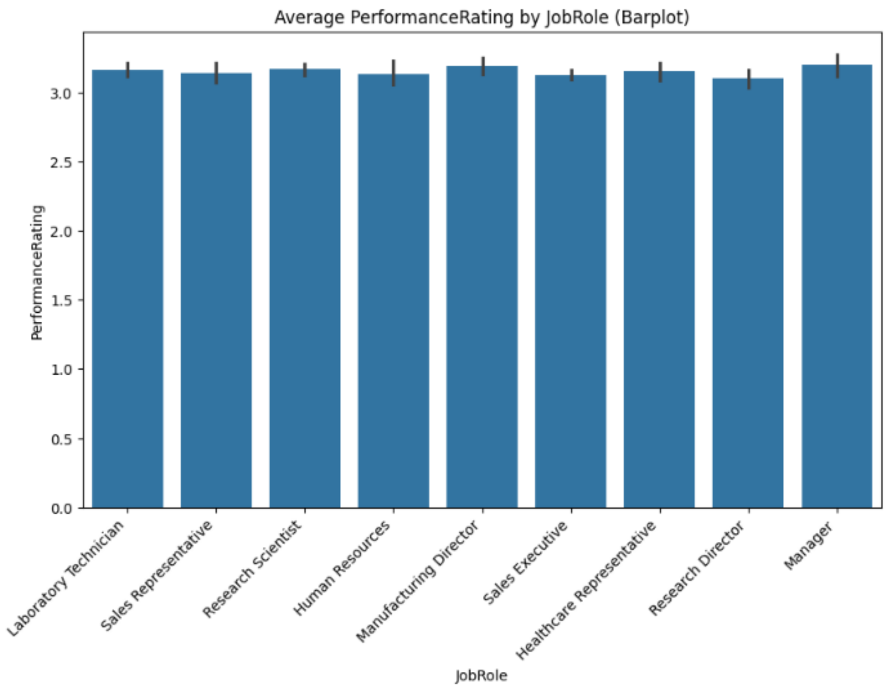
Attrition Count by Department: Attrition is highest in Research & Development and next in Sales, with Human Resources showing very few cases. This mirrors department sizes, so use rates (attritions per headcount) for fair comparison, but the visual highlights where most absolute exits occur.

MonthlyIncome distribution (Histogram + Boxplot)



MonthlyIncome Distribution and Boxplot: Salaries are right-skewed—many employees earn in the lower to mid ranges, while a smaller group earns very high incomes. The boxplot shows a wide IQR and numerous high-end outliers, confirming large dispersion and a long upper tail.

PerformanceRating vs JobRole



Average PerformanceRating by JobRole: Mean ratings cluster tightly around 3.1–3.2 for every role, with small error bars. This suggests consistent performance evaluations across roles and limited variance, implying standardized appraisal outcomes rather than role-driven differences.

Dataset Observation

Workforce size: 1,470 employees; 3 departments (R&D 961, Sales 446, HR 63).
Salaries (MonthlyIncome): right-skewed with high outliers (~20k); wide spread within departments; leadership roles (Manager, Research Director) earn the most.
Age: broad spread from 18 upward; strong mid-career concentration; meaningful >45 segment extracted for analysis.
Attrition: highest counts in R&D, then Sales; HR minimal. Compare rates in future for fairness.
Performance: average ratings tightly clustered ~3.1–3.2 across roles, indicating consistent evaluations.
JobRole insights: IC roles (Lab Technician, Research Scientist, Sales Rep)
sit in lower pay bands; managerial/dir roles lead compensation.

Conclusion

- Overall, the organization is mid-sized (1,470 employees) with an R&D-centric structure, a sizable Sales team, and a lean HR function. Compensation is right-skewed with wide dispersion; leadership roles command the highest pay while several individual-contributor roles sit in lower bands. Performance ratings are remarkably consistent across roles (~3.1–3.2), and attrition counts align with headcount, appearing highest in R&D and Sales. The dataset is clean and analysis-ready; adding location and bonus fields and using rate-based metrics (e.g., attrition per headcount) will unlock deeper, fairer comparisons and sharper workforce decisions.